

04P02302



83

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

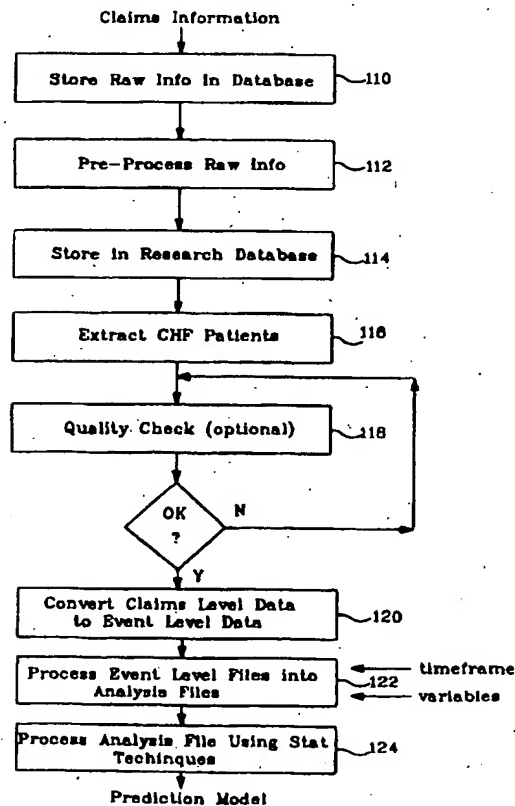
## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> : <b>G01N 33/48, A61B 5/00, G06F 17/60</b>		<b>A1</b>	(11) International Publication Number: <b>WO 97/28445</b>
			(43) International Publication Date: 7 August 1997 (07.08.97)
(21) International Application Number: <b>PCT/US97/01829</b> (22) International Filing Date: <b>3 February 1997 (03.02.97)</b> (30) Priority Data: 60/011,772                      2 February 1996 (02.02.96)                      US (71) Applicant: <b>SMITHKLINE BEECHAM CORPORATION</b> [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). (72) Inventor: <b>WONG, Bruce, Jan, On</b> ; 1254 Gulph Creek Road, Radnor, PA 19087 (US). (74) Agents: <b>KANAGY, James, M. et al.</b> ; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).			(81) Designated States: AU, CN, JP, KR, MX, NZ, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  Published With international search report.

(54) Title: METHOD AND SYSTEM FOR IDENTIFYING PATIENTS AT RISK FOR AN ADVERSE HEALTH OUTCOME

## (57) Abstract

A computer-implemented technique, including database processing, is used for identifying the existence of at risk patients in a claims database (114). Claims information for a group of depression patients is extracted from the database (116). A set of events is defined using the extracted information. This information and the set of events are processed to create event level information (120). A time window is defined for providing a timeframe from which to judge whether events should be considered in subsequent processing and, a set of variables is defined as being potential predictors of adverse health outcomes. The event level information, using the time window and the set of variables, is processed to generate an analysis file (122). Statistical analysis is performed on the analysis file (124) to generate a prediction model which is a function of a subset of the set of variables.



BEST AVAILABLE COPY

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

## METHOD AND SYSTEM FOR IDENTIFYING PATIENTS AT RISK FOR AN ADVERSE HEALTH OUTCOME

### BACKGROUND OF THE INVENTION

This invention relates to a proces for identifying patients with a specified  
5 disease who are at risk for a near-term high-cost clinical outcome associated with  
that disease. The technique may also include providing early notice to a medical  
services provider of the at-risk patient, or to the patient directly. It may also  
include providing one or more interventions which can modify said clinical event  
and risk of high-cost clinical outcome. More particularly, it relates to the  
10 identification of patients diagnosed with depression and having a high risk of  
adverse health outcomes by using various database processing techniques.

With regards to depression, it is one of the most common treatable  
conditions affecting our society. In fact, depression occurs at levels comparable  
to angina and coronary artery disease. The point prevalence for major depression  
15 in western industrialized nations is 2.3 to 3.2% for men and 4.5 to 9.3% for  
women. The life time risk for depression is 7 to 12% for men and 20 to 25% for  
women. These statistics reflect the substantial burden depression has on society.

The economic burden of depression, however, is more difficult to  
quantify. Some estimates show that depression accounts for approximately 1/3 of  
20 the direct costs of all mental illnesses (\$67 billion in 1990). Of the depression  
related costs, approximately 2/3 are related to direct medical expenditure.  
Although estimates of the economic cost of depression vary, they were  
conservatively estimated at \$16.3 billion in 1980 of which approximately 2/3 were  
direct medical costs.

25 Depression is widely perceived as an essentially self-limiting condition,  
where a background of good functioning is punctuated by brief periods of illness  
and subsequent recovery. Over 50% of patients have recurrent episodes of  
depression. Treatment can then be viewed as being of an episodic nature with  
management of individual episodes.

30 The current Practice Guideline For Major Depressive Disorder in Adults  
published by the American Psychiatric Association (APA) in 1993 describe  
various means of diagnosing and treating depression and is herein incorporated  
by reference for its teachings about depression diagnosis and treatment. Other  
literature also exists, for example, literature published by the Agency of Health  
35 Care Policy Reasearch (AHCPR), which describes the illness, its symptoms and  
means of diagnosis and treatment. These materials can be used in the  
interevention step of this invention as a means for modifying the habits or course

of medical treatment in a way which can prevent or reduce the identified high-cost clinical outcome predicted to possibly occur using the methods describe herein. In addition, custom programs and materials can be developed based on this Practice Guideline and other medical and clinical information. These custom programs and materials can be based in part or in whole on the identified predictive events and the relative weightings given to each event.

To date, the treatment of depression has been on an individual basis. Numerous reasons exist, however, for the cessation of individual treatment regimes including all of those factors which ordinarily input to a "cost-benefit" analysis at an individual level (likelihood of further improvement, severity of illness, medication side-effects, etc.).

Thus, it appears that, in view of the overall burden depression creates for society - particularly the financial burden - alternative means of treating depression need to be explored. For example, evidence exists in support of the efficacy of maintenance chronic therapy. Under this theory, the clinical goal would be the maintenance of euthymia, not repetitive treatment of recurrent episodes which may contribute to a deteriorating lifetime course.

Under this theory, however, figures appear to indicate that it may only be viable to treat a portion of the depression-diagnosed population in this way, perhaps, with targeted interventions at subgroups at risk of adverse outcomes (in particular, recurrence). There is, therefore, a need to be able to accurately and effectively identify subgroups of the depression population at high risk of adverse health outcomes.

#### SUMMARY OF THE INVENTION

In its broadest embodiment, this invention involves a computer-implemented method for identifying at least one patient with a specified disease who is at risk for a near-term high-cost clinical outcome associated with said disease and at least one clinical event predictive of at least one high-cost clinical outcome attributed to said patient, and providing early notice to a medical services provider of said at-risk patient or said patient, and optionally providing to the provider or patient one or more interventions which can modify said clinical event and risk of high-cost clinical outcome, which method comprises:

- i) collecting relevant data on patients who have a current diagnosis of the specified disease from one or more data sources;
- ii) entering or merging said data into a single electronic data file;
- iii) optionally cleaning up said data file by removing extraneous data and associating patient data with the correct patient identifier;

iv) identifying a clinically relevant, appropriate time period;

v) creating a patient data file identifying, within the clinically relevant time period, for each patient the presence and frequency of at least one predetermined clinical event predictive of risk for a near-term high-cost clinical outcome relevant to the specified disease;

vi) stratifying the identified patients into risk groups based on the found predetermined clinical event by processing through a computer the patient data file, using a second pre-existing data file, wherein the second pre-determined data is created using a statistical procedure and a data file generated by the steps of:

a) processing, based on predetermined criteria, patient information in at least one clinically relevant data source to extract information for a group of patients with the specified disease;

b) defining, using the information available in said data source, a set of potentially predictive events relevant to a high cost clinical outcome in patients with the specified disease;

c) defining a time-window from which to judge whether events should be considered in subsequent processing;

d) creating files for predicting clinical outcomes containing relevant data rearranged to represent the set of potentially predictive events; and

e) running a statistical procedure to identify a set of predictive events which in combination show a statistically significant association with a high-cost clinical outcome;

vii) notifying the healthcare provider of said patient or patient that said outcome is likely to occur in the near term, and

viii) optionally providing to the health care provider or patient, interventions which modify the high-cost clinical outcome or reduce its severity, which interventions are optionally derived from the one or more of the set of predictive clinical events which in combination show a statistically significant association with a high-cost clinical outcome.

The present invention also involves a computer-implemented method for generating a model to identify at risk patients diagnosed with depression, information about patients existing in a claims database, said method comprising the steps of 1) processing, based on predetermined criteria, the patient information in the claims database to find and extract claims information for a group of depression patients; 2) defining, using the information available in the claims database, events relevant to depression; 3) processing the extracted claims

information and the defined events to create files containing event level information; 4) defining a time window for providing a timeframe from which to judge whether events should be considered in subsequent processing; 5) defining a set of variables as potential predictors; 6) processing the event level information, using the time window and the set of variables, to generate an analysis file; and 7) performing statistical analysis on the analysis file to generate a prediction model, said prediction model being a function of a subset of the set of variables.

Another aspect of the present invention involves a computer-implemented method for identifying, using the generated model, at risk patients diagnosed with depression, said method comprising the additional step of applying the prediction model to a processed claims database to identify and output a file listing the likelihood of each patient having an adverse health outcome.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The invention is best understood from the following detailed description when read in connection with the accompanying drawing, in which:

Figure 1A is a high-level flowchart illustrating an exemplary overall process of the present invention.

Figure 1B is a high-level flowchart illustrating an exemplary process of the application of the present invention.

Figure 2 is a high-level block diagram illustrating three exemplary sources of information suitable for use with the present invention.

Figure 3 is a data structure diagram which shows an exemplary format in which the information from the sources of Figure 2 are stored in a research database.

Figure 4 is a data structure diagram which shows an exemplary format for an event level file generated during the process shown in Figure 1.

Figure 5 is a data structure diagram which shows an exemplary format for an analysis file generated, in part, from the event level file shown in Figure 4 and during the process shown in Figure 1.

Figure 6A is a time-line diagram which shows a first exemplary time window scheme suitable for use in processing the data from the event level files shown in Figure 4.

Figure 6B is a time-line diagram which shows a second exemplary time window scheme suitable for use in processing the data from the event level files shown in Figure 4.

Figure 7A is a table which shows experimental results using a hospitalization (HL) indicator with the Scheme 1 shown in Figure 6A.

Figure 7B is a table which shows experimental results using a High Cost indicator with the Scheme 1 shown in Figure 6A.

## DETAILED DESCRIPTION OF THE INVENTION

### Overview

5        This invention involves a computer-implemented method for identifying at least one patient with a specified disease who is at risk for a near-term high-cost clinical outcome associated with said disease and at least one clinical event  
10        predictive of at least one high-cost clinical outcome attributed to said patient, and providing early notice to a medical services provider of said at-risk patient or said patient, and optionally providing to the provider or patient one or more  
15        interventions which can modify said clinical event and risk of high-cost clinical outcome. It also involves a means for controlling health care costs by notifying the health care provider, or the patient, of the existence of risk factors, the likelihood of a near-term event, meaning within 1 to 6 months or so. In addition  
20        information about the risk and how to avoid it or reduce the risk or the cost of the near-term event may be provided to the health care provider or to the patient. Sources for such so-called interventions are set out above herein.

      Unless otherwise defined herein, a clinical event is inclusive of all activities related to health and health care which may have a impact, usually  
20        negative, on the health of an individual presently or in the future. This term is inclusive of genetically determined traits such as gender. It includes a life-style or job or work environment events which might impact on health. It also includes the likes of diagnoses; treatments, such as direct physical interventions or prescribed therapies such as drugs. This list is intended only to illustrate, not limit, what may  
25        fall within the perview of a clinical event.

      A clinical outcome is future event which is a consequence of an existing clinical event or one which has happened in the recent past.

      A time-window, as that term is used herein, is disease and clinical event dependent. This window has been illustrated for depression, infra; the time-  
30        window is essentially the same for generating a file showing statistical significance between clinical event and for generating the stratified grouping of patients with near-term high-risk clinical outcomes. But these two windows need not be the same, though the window for obtaining the predictive clinical events will almost always be at least as long as the window for the file from which is generated the  
35        stratified patients groupings. It will be appreciated that this time-window may be quite long for diseases which have a long latency period, or which develop over many years as a result of episodic or continued clinical events. For example an

analysis of predictive clinical events to determine which patient may be at risk in the near term for Alzheimer disease of either type may need to take into account historical data going back 5, 10, 15 or more years. While it is expected that most diseases will be amenable of analysis using data going back 1 to 5 years, the invention is not to be so limited.

More specifically, the present invention is designed to identify, in a predetermined population of depression patients, those patients at high risk of adverse health outcomes, also called clinical outcomes. The identification of this high risk subgroup being an initial stage in attempts, e.g., targeted interventions, to prevent and/or improve their health outcome.

Initially, one or more sources of information are required which allows for the identification of an initial population of depression patients. Examples of sources include health care providers such as doctors, hospitals and pharmacies which all keep records for their patients. The individual records for each of these providers, however, may be scattered, difficult to access, and/or have many different formats.

On the other hand, a more comprehensive source containing this type of information exists in the health care claims records of any given benefits provider.

Turning to the figures, Figure 1A is a high-level flowchart illustrating an exemplary overall process of the present invention. As illustrated in Figure 1, the "raw" claims information is received and stored in a database (e.g., DB2 format) represented by block 110. In the world of claims processing, before this database of "raw" information can be useful, some pre-processing, step 112, is generally performed which may include rejecting claims, reconciling multiple claims and so on. The output of this preprocessing step, represented by block 114, is a "cleaner" database now stored, in the exemplary embodiment, in SAS format.

SAS is a well known format and software package produced by SAS Institute, Inc. of Cary, North Carolina. It should be noted that other data processing and storage formats, as appreciated by those skilled in the art, could be used in the storage and processing of data.

It should also be noted that SAS formats, programming techniques and functions are more fully described in the SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 and 2, 1990 and the SAS Language: Reference, Version 6, First Edition, 1990 which are both herein incorporated by reference for their teachings regarding the SAS language, SAS programming, functions and formats.



Moreover, the SAS routines used for processing information as part of the present invention are used for computational operations, executed on a computer and stored on a storage medium such as magnetic tape, disk, CD ROM or other suitable medium for purposes of storage and/or transportability. The stored software can then be used for running a computer.

The claims records of the benefits provider, although containing important information, may not be organized in a manner for efficient analysis. Thus, the next step is to perform another processing step (e.g., screening for depression patients, age, etc.), represented by block 116, to transform the "raw" data into a more appropriate and useful database. That is, the output data from the processing (i.e., extraction) step is a subset of the "raw" information and represents an initial universe of depression patients upon which further processing is performed.

A next step, which is optional, is to perform a "quality check" on the initial universe of depression patients. This step is somewhat subjective. This processing step, represented by block 118, using intermediate output files, performs a refinement of the extracted information by, for example, checking to see if an imbalance exists in the extracted information such as all claims are from individuals over 60 years of age or all claims are from men. This step, essentially a common sense check, can be performed as many times as necessary to ensure the integrity of the database data. At this point, the database data exists at the claim level.

The information existing at the claim level provides various information in the form of raw data elements. From the claims level data, the next processing step, represented by block 120, creates new files (e.g., primary file 1 and primary file 2) by reformatting the information into an event level format.

Before this occurs, a set of events (e.g., doctor visit for depression) relevant to depression are defined using a combination of both the raw data elements available from the claims information and clinical knowledge about depression. With these events defined, the claims level information is used to create new files based on events rather than claims. Having the information in an event level format is an important aspect of the present invention in that, among other things, it allows for added flexibility in subsequent analysis.

As depicted by block 122, further processing is performed on the event level data to generate an analysis file. In particular, the processing is performed using input information representative of a sliding time window and a plurality of variables. The time window input limits the time periods in which the events

from the primary files are considered. That is to say, the time window is used to identify an analysis region and a prediction region where activity in the analysis region is used to predict some predetermined outcome in the prediction region. The selection of variables, both dependent and independent, for analysis, is an important step impacting the accuracy of the final prediction model. The dependent variables are representative of the desired result (i.e., an adverse health outcome to be predicted); whereas, the independent variables are representative of predictors. This processing step, step 122, can be easily re-programmed, via the input parameters, for various time window adjustments as well as various variable modifications. The analysis file generated at this step is a member level file which means it is broken down by member.

With the analysis file in hand, a model or technique for identifying high risk subgroups is determined. That is, as represented by step 124, the analysis file is used to develop an identification technique represented by an equation incorporating a subset of the initial variables programmed into the above-mentioned processing step. The resulting subset are those variables which best reflect a correlation to adverse health outcomes, consequently, resulting in substantial use of health care resources (e.g., funds). It should be noted that the determination of the initial as well as the final variables is an important aspect of present invention as the variables may significantly impact the accuracy of the identification of the subgroup.

The above model for identification can be developed, step 124, in various ways using statistical techniques. The technique used in the exemplary embodiment of the present invention for generating the model is multiple logistic regression.

Figure 1B is a high-level flowchart illustrating an exemplary process of the application of the present invention. Having developed the model, as shown in Figure 1A, it can then be applied to updated claims data, step 132, or to other databases of depression patients (e.g., claims information for other benefits providers), in order to identify at risk patients diagnosed with depression, step 134, allowing for various types of targeted intervention to maximize the effective allocation of health care resources.

Several examples are set out hereafter to illustrate the invention. They are given solely for purposes of exemplification and are not intended to limit the invention in any manner or any fashion.

### Exemplary Embodiment of the Invention

Although the present invention is illustrated and described below with respect to specific examples of a method and system for identifying depression patients at high risk for adverse health outcomes, the invention is not intended to be limited to the details shown. Rather, various modifications may be made in the details within the scope and range of equivalents of the claims and without departing from the spirit of the invention.

As mentioned, the present invention is designed to identify patients with depression at high risk of adverse health outcomes. The identification of this high risk subgroup being the first step in being able to try different treatment techniques (e.g., targeted interventions).

Initially, a source of information is required which allows for the identification of a population of depression patients. A comprehensive source containing this type of information exists in the health care claims records of many benefit providers. As is known, claims for drugs, doctors and hospitals are received and processed for payment/reimbursement. In the exemplary embodiment of the present invention, this claims information is entered into a DB2 database on a benefits provider's computer system (not shown).

Figure 2 is a high-level block diagram illustrating three exemplary sources of information suitable for use with the present invention. As illustrated in Figure 2, the claims information of such a provider would typically include three sources: pharmacy claims (Rx) 210, doctor (DR) claims 212, and hospital (HL) claims 214. As listed on the blocks representing the claims information, many types of information would be available from the respective claims including drug codes, physician's names, diagnosis codes, procedures, various dates and other important information. Much of this information is referenced using codes, such as drug codes, procedure codes and illness codes. Appendices I-VI provide listings of various codes used with the present invention. These codes were selected for processing purpose of the present invention from a voluminous source of codes and, as will be appreciated by those skilled in the art, may be modified to include/exclude codes deemed more/less useful at the various stages of processing.

The DB2 database represents a source of "raw" data elements which require processing. A first step in processing this raw data is to perform data integrity checks (e.g., rejected or reconciled claims). Subsequently, the data is routinely download into a "research" database. The research database is a claims level database in SAS format.

Exemplary formats, for each of the Rx, DR and HL claims, of the records contained in the research Database, are shown in Figure 3. As shown in Figure 3, claims are listed from claim 1 to claim x and the appropriate information, for the particular service provider (e.g., Rx) being claimed, is also presented.

5        Once in SAS format, SAS procedures process the information to 1) extract patients with depression (step 116), 2) process the claims level information into event level information (step 120), 3) using predetermined variables and timeframe schemes, generate analysis files for analysis purposes (step 122) and 4) create a prediction model as a function of those variables most reflective of the  
10       correlation to an adverse health outcome (step 124).

It should be mentioned that, from a statistical perspective, an important consideration in developing prediction models from datasets is sample size. To maximize the integrity of the prediction model, sample size is an important factor. Prevalence of depression is reported to be approximately 5%, however, sample  
15       sizes required to determine prediction equations depend on the magnitude of association between variables. As these associations are unknown, all patients within any individual plan are initially included.

The first step, extracting patients with depression (step 116), uses various parameters to define which patients qualify for the overall initial universe of  
20       depression patients to be considered.

For example, in the exemplary embodiment of the present invention, only patients having a continuous enrollment with the benefits provider of 12 months or longer and having a claim for depression or treatment with anti-depressant medication are eligible. Of course, these criteria are exemplary and could be  
25       modified such that 24 months or 6 months of enrollment is satisfactory or that an individual must be 18 years of age. In the exemplary embodiment of the present invention, the claims extraction step, step 116, extracts all claims data for patients with either an appropriate code for depression (see Appendix I) or for treatment with an antidepressant drug (see Appendix III).

30       It should be noted that in the health care industry various codes are used in claims information for indicating which procedures, treatments, diagnoses, drugs, etc. are being claimed. For the exemplary embodiment of the present invention, the selected codes are shown in Appendices I-VI. These codes were found in Physician's Current Procedural Terminology (CPT), American Medical  
35       Association (1995) and St. Anthony's ICD-9-CM Code Book (1994) which are both herein incorporated by reference for their teaching of codes and sources of codes. As will be appreciated by those skilled in the art, any set of codes,

representative of the various procedures, treatments, diagnosis, drugs, etc. relevant for use with the present invention would suffice. Reference to such codes occurs throughout this specification.

Subsequent to the claim extraction step, the claim adjustment and integrity checks are optionally performed, step 118. To do so, from the dataset defined above, intermediate output files are generated which contain sets of frequency counts for processing purposes. In the exemplary embodiment of the present invention, intermediate output files for the following characteristics are generated for review:

- 10 a. frequency counts of unique members by sex, age groups (0-9, 10-19...) and enrollment duration by months including:
  - i) Tables showing count of members by sex, ii) Table showing count of members within age groups, iii) Table of counts of age groups broken down by sex, iv) Table of enrollment duration by months i.e., 1 month to  
15 maximum number of months possible.
- b. frequency counts of ICD codes for depression (Appendix I), i.e., number of members having at least one hit with each of the ICD codes in Appendix I any level ii) as first code.
- c. frequency counts of anti-depressant drugs (Appendix II):
  - 20 i) number of members who have at least one claim for each of the drugs in Appendix III.
- d. count of members who became eligible for processing due to ICD code only, by drug only, and by both ICD code and drug.
- e. frequency counts of numbers of all claims within each file (HL, DR, Rx) by member.  
25
- f. frequency counts of ICD codes (use only the first 3 digits of ICD codes) of any nature in DR (any position) and HL files - at least the top 10 with frequency of each. i.e., 2 tables one each for DR and HL files.
- 30 g. frequency counts of hospitalizations by calendar month. Counting calendar month backward from last month of eligibility or data availability. The last month for which data is available will be month 1, the penultimate month with be month 2 etc.
- h. frequency counts of procedures related to depression (CPT codes, Appendix II).
- 35 i. frequency counts of all CPT codes (to the level of the first 3 code digits) - at least the top 10.

The above frequency counts for use in performing preliminary evaluations as to the integrity of the data is exemplary and could be modified to include/exclude parameters which are shown to be more/less useful.

With this information, a "quality check" is performed on the initial  
5 universe of depression patients to make sure that the final results, i.e., prediction model, is not unreasonably skewed due to bogus input information. This processing step, block 118, using intermediate output files, allows for a refinement of the extracted information by, for example, checking to see if an  
10 imbalance exists in the extracted information such as all claims are from individuals over 60 years of age, all claims are from men, or other data imbalances which would otherwise taint the integrity of a prediction model. Step 118, in the exemplary embodiment, is performed manually by viewing the intermediate output files. It is contemplated, however, that using various threshold values, the frequency counts can be automatically scanned for a  
15 potential imbalance.

Having now extracted and refined the claims level information according to various predetermined criteria deemed relevant for subsequent processing purposes, the information is converted into an event level format.

To provide processing flexibility, particularly in assigning time windows  
20 for analysis, the above-mentioned second step (i.e., converting the claims level information into event level information, step 122) is employed to generate two primary data files from which an analysis file can be created.

In the exemplary embodiment of the present invention, primary data file 1 is a member level file and contains all data of a static nature (i.e., not time  
25 sensitive) such as 1) Member Key, 2) Date of birth, 3) Gender, 4) First available date of enrollment (i.e., start of dataset (1/1/92) or enrollment date), 5) End date of enrollment (i.e., end of dataset or last date of enrollment), 6) Date of first depression event (first prescription for antidepressant or depression hospitalization), 7) Date of last hospitalization, 8) Number of records in events  
30 file (primary file 2), and 9) Mode of entry into the dataset (e.g., i) Anti-depressant drug only, ii) Depression diagnosis only, iii) Both anti-depressant drug and depression diagnosis).

Primary data file 2 is an events level file with a record for each event ordered by member and the chronological date of the event, in the present  
35 invention, presented in descending order of event date.

It should be noted that an event, sometimes referred to as an episode, is an occurrence which, based on clinical knowledge, is deemed relevant to depression.

Having knowledge of what raw data elements are available from the claims, a set of events is defined directly or indirectly from the data elements where events can be based on an individual data element, combination of data elements or derived from individual or multiple data elements.

5           Figure 4 is an exemplary list of events and format for primary file 2 (an event level file). As shown in Figure 4, the entries provided include:

1. Hospitalization for depression
    - a. Any hospital claim identified by hospital site code.
    - b. Having a from and through duration of at least 1 day.
    - 10       c. Having ICD 9 code.
    - d. Depression ICD 9 code occurring at any position.
    - e. Illness indicator (Appendix V) 1 = major illness, 2 = suicide, 3 = major illness and suicide. 0 = everything else.
  2. Emergency room for depression
    - 15       a. Emergency room visit identified by emergency room site code.
    - b. Having ICD 9 code (see Appendix I).
  3. Doctor (non-hospital) visit for depression
    - a. Any doctor claim.
    - b. Having ICD 9 code (see Appendix I).
    - 20       c. Category : Psychiatrist = 1, all others = 0.
  4. Prescription for SSRI
    - a. SSRI (selective serotonin re-uptake inhibitors) therapeutic class 5.5.1.3.
    - b. Cost = 0 if generated from a hospital admission.
    - 25       c. Category indicator = blank
  5. Prescription for (Tricyclic antidepressants) TCA or (Monoamine Oxidase Inhibitors ) MAOI
    - 30       a. Therapeutic classes 5.5.1.1 (tertiary amines), 5.5.1.2 (secondary amines), 5.5.1.4 (Monoamine Oxidase inhibitors). AND 5.5.2
    - b. Cost = 0 if generated by a hospital admission
    - c. Category indicator = therapeutic class 1 = 5.5.1.1, 2 = 5.5.1.2, 3 = 5.5.1.4, 4 = 5.5.2
  6. Prescription for other neuroactive drug (From Rx file)
  - 35       7. Procedure for depression (from DR or HL files)
- Category:
- CPT codes or ICD procedure

- 0 = Psychotherapy      All CPT and ICD codes in  
Appendix II not listed  
below.
- 5      1 = Diagnostic      90801, 90820, 90825, 90830,  
90862  
94.0x, 94.1x, 94.21, 99.22,  
94.23
- 2 = Shock therapy      890870, 908712  
94.24, 94.26, 94.27
- 10      For this entry, costs are assigned to the doctor visit or hospitalization in  
which the procedure occurred.
8. Hospitalization not for depression
- It should be noted that items under entry 8 could have been performed for  
a condition other than depression although these patients got into the cohort by  
15      virtue of receiving a depression diagnosis or receiving and antidepressant at some  
time making it likely these procedures were for depression.
- a. All hospitalization having from and through dates of at least one  
day duration.
- b. Major illness ICD 9 codes (see Appendix V).
- 20      c. Category as in 1 above (1 = major, 2 = suicide, 3 = both, 0 = all  
others)
- Counts for entries 9-13 are aggregated for each month. The date is that for  
the first occurrence of the identified events. In the number field, the number of  
identified events occurring in that month are summed.
- 25      9. Emergency room not for depression
- a. Emergency room visit identified by Emergency room
10. Doctor (outpatient) visit not for depression
- a. Any doctor visit.
- b. Excluding visit with a depression diagnosis (Appendix I) i.e.,  
30      not in 3/above.
11. Prescription for possibly related drugs  
Drugs identified in Appendix IV
12. Prescription for all other (non-depression) drugs  
All drugs not included in Appendices III or IV.
- 35      13. Procedure not for depression (from Dr and HL files)
- a. Category indicator 1 = major procedures, 2 = minor procedure  
(see Appendix IV).



After generating the two primary files using the above described instructions, corresponding to step 120 of Figure 1, further processing is performed on the event level data to generate an analysis file, step 122. An exemplary format for the analysis file is shown in Figure 5. As shown, the format of the analysis file includes a list of members in a first column of a table. Across the top of the table is a list of variables, described in detail below. And, the body of the table provides indications as to a member's relation to a listed variable.

In particular, the processing from the primary files to the analysis files includes an algorithm defined, in part, by a time window and a plurality of variables. The algorithm can be re-programmed for various time window adjustments as well as variable modifications. The analysis file generated at this step is a member level file (i.e., organized with respect to members). The main analysis files are member level files derived from the information in the primary files.

Each main analysis file is created to take into account a single reference time window of censored events and prediction window of interest for that file. Each new time window applied to the data, in the exemplary embodiment, requires another main analysis file.

To generate the analysis file, a time window scheme, along with a plurality of variables, is applied to the event level data.

Discussing the variables first, included in the processing are both independent and dependent variables. The independent variables basically represent potential predictors of the adverse health outcomes; whereas, the dependent variables basically represent the adverse health outcome to be predicted.

To determine exemplary independent variables for step 122, as many of the original data elements as possible are used, assuming nothing about depression. Then, based on clinical knowledge, additional variables are created. Furthermore, combinations of the data elements and/or variables, based on clinical knowledge, are used as variables. Finally, some variables may be created and used based on their potential utility as a leverage point in disease management.

It should be noted that, for purposes of a cost hierarchy, the following rules were used in the exemplary embodiment of the present invention.

1. Only hospitalizations for depression can spawn other events.
2. Hospital costs include all Rx, procedure, physician charges.
3. Hospital visits can generate Rx and procedure events with costs set to zero (included in hospital cost).

4. Hospital visits cannot generate separate doctor visit events.

In the exemplary embodiment of the present invention, the plurality of variables currently used by step 122 in the SAS routine for generating an analysis file from an event level file are shown below in Table 1. In Table I, although the abbreviations should be self-evident, by way of example, some abbreviations are as follows: "DEP" means depression, "HL" means hospitalization, "#" means number, "MOS" means months, "OTH" means other, "ER" means emergency room, "RX" means prescription, "SUP" means supply, "PROCS" means procedure, and "TOT" means total.

10 Table 1

1. "DEPENDENT CODE"
2. "DEPRESSION HL INDICATOR"
3. "# OF MOS AVAILABLE FOR ANALYSIS"
4. "AGE AT TIME OF CUTOFF"
- 15 5. "FEMALE INDICATOR"
6. "TOTAL COST DURING ANALYSIS PERIOD"
7. "# OF DEPRESSION DRUG CLASS SWITCHES"
8. "DEPRESSION DRUGS DAYS SUPPLY"
9. "# DEP HLS"
- 20 10. "# DEP OTH HLS"
11. "# DEP HLS AND MAJOR ILLNESS"
12. "# DEP HLS AND SUICIDE"
13. "# DEP HLS AND MAJOR ILLNESS AND SUICIDE"
14. "# DEP HLS AND DEP RELATED CODE"
- 25 15. "# DEP HL LENGTH OF STAY"
16. "# DEP ER VISITS"
17. "# DEP DR VISITS"
18. "# DR/PSYCHIATRIST VISITS"
19. "# RX FOR SSRI"
- 30 20. "# DAYS SUP OF SSRI"
21. "# RX FOR TCA"
22. "# RX TCA: TERTIARY"
23. "# RX TCA: SECONDARY"
24. "# RX TCA: MONO OXI INHIBITORS"
- 35 25. "# RX TCA: ALL OTHER TYPE"
26. "# DAYS SUP OF TCA"
27. "# DAYS SUP OF NEUROACTIVE (NA)"

28."# DYS SUP OF NA: ANXIOLYTICS AND SEDATIVE"  
29."# DAYS SUP OF NEUROACTIVE: ALL OTHER"  
30."# DEPRESSION PROCS"  
31."# DEP PSYCHOTHERAPY PROCS"  
5 32."# DEP DIAG PROCS"  
33."# DEP SHOCK THERAPY PROCS"  
34."# OTH HOSPITALIZATIONS"  
35."# OTH ALL OTH HLS "  
36."# OTH HLS AND MAJOR ILLNESS"  
10 37."# OTH HLS AND SUICIDE"  
38."# OTH HLS AND MAJOR ILLNESS AND SUICIDE"  
39."# OTH HLS AND DEP RELATED CODE"  
40."# OTH HL LENGTH OF STAY"  
41."# OTH ERS"  
15 42."# OTH DR VISITS"  
43."# RX FOR RELATED DRUGS"  
44."# DAYS SUP RX FOR RELATED DRUGS"  
45."# RX FOR ALL OTHER"  
46."# DAYS SUP RX FOR ALL OTHER"  
20 47."# PROCS NOT FOR DEP"  
48."# PROCS FOR MAJOR ILLNESS"  
49."# PROCS FOR MINOR ILLNESS"  
50."% DEP HL COST OF TOT COST"  
51."% DEP ER COST OF TOT COST"  
25 52."% SSRI COST OF TOT COST"  
53."% TCA COST OF TOT COST"  
54."% NEUROACT COST OF TOT COST"  
55."% OTH HL COST OF TOT COST"  
56."% OTH ER COST OF TOT COST"  
30 57."% OTH DR COST OF TOT COST"  
58."% OTH RELATED RX COST OF TOT COST"  
59."% OTH ALL OTHER RX COST OF TOT COST"  
60."# COST OF DEP RELATED EVENTS"  
61."# COST OF OTH RELATED EVENTS"  
35 62."# COST DEP DRUGS FROM ALL DRUG COSTS"  
63."# COST OTH DRUGS FROM ALL DRUG COSTS"  
64."# COST DEP DRUGS FROM ALL COSTS"

- 65. "# COST OTH DRUGS FROM ALL COSTS"
- 66. "# SSRI COST FROM DEP DRUGS COSTS"
- 67. "% TCA COST FROM DEP DRUGS COSTS"
- 68. "% NEUROACTIVE COST FROM DEP DRUGS COSTS"
- 5 69. "DEP HL IN LAST 12 MOS INDICATOR"
- 70. "DEP ER IN LAST 12 MOS INDICATOR"
- 71. "MOS BETWEEN 1ST AND LAST EVENT"
- 72. "MOS SINCE FIRST DEP EVENT"
- 73. "MOS SINCE LAST FIRST DEP EVENT"
- 10 74. "MOS OF DATA USED FOR ANALYSIS"
- 75. "DEP RX COMPLIANCE MEASURE"
- 76. "# DEP HL BY GENDER INTERACTION"
- 77. "# DEP ER BY GENDER INTERACTION"
- 78. "# DEP DR VISITS BY GENDER INTERACTION"
- 15 79. "# RX SSRI BY GENDER INFORMATION"
- 80. "# RX TCA BY GENDER INTERACTION"
- 81. "# RX NEUROACTIVE BY GENDER INTERACTION"
- 82. "# DEP PROCS BY GENDER INTERACTION"
- 83. "# OF UNIQUE GENERALIST DRS USED"
- 20 84. "# OF UNIQUE PSYCHIATRISTS USED"

Turning to the dependent variables, potential dependent variables, for example, contemplated for use with the present invention as results to be predicted include:

1. Hospital (HL) admission or emergency room (ER) visit for depression.  
 25 This is a dichotomous variable which is referred to as the HL (or ER) indicator such that HL (or ER) = 1 if an admission or ER visit occurred, otherwise the indicator equals 0.
2. Highest 10% of resource utilization measured in dollars. Resources counted from time of cost in the top 10% of the first depression diagnosis or  
 30 receipt of first antidepressant (in the record) + 1, 3 and 6 months - separate analyses for each time period. Again, this is a dichotomous variable referred to as the High Cost indicator such that if patient in top 10%, Host Cost = 1, otherwise High Cost = 0. The High Cost indicator, in the exemplary embodiment, could also be defined as the distribution of total cost per member (PMPM) in the prediction  
 35 region (B to C) is used to define this variable. The High Cost indicator is set to 1 for the 10% of members with the highest PMPM in the Total Cost distribution and set to 0 for all others.

3. Any hospital admission for attempted suicide - identified by claim related to any of the ICD 9 codes 300.9 or 800-999. As those of ordinary skill in the art will appreciate, using attempted suicide as a dependent variable may only provide useful results if there exists a sufficient number of occurrences to do so.

5 Although only three dependent variables are listed above, as those of ordinary skill in the art will appreciate, other known or yet unknown variables may also suitably serve as a dependent variable within the scope of the present invention.

Turning to the time window aspect of the generation of the analysis file, it should be noted that there is one analysis record for each selected member.

10 In the present invention, several schemes, as described below, have been developed for defining prediction zones and censoring data to create the analysis file. That is, a time window basically defines a prediction zone or region and an analysis region where the analysis region in where activity is used to predict something in the prediction zone. Additional time window schemes may also

15 adequately serve the present invention.

For purposes of explanation, the time that the claims history covers is referred to as the time window that starts at point 'A' and ends at point 'C'. The time interval is divided into analysis and prediction regions by point 'B' such that

20  $A < B < C$ .

By way of example, Jane Doe's analysis record is based on claims from 1/1/91 through 6/30/93. Therefore,  $A=1/1/91$ ,  $C=6/30/93$  and B can be selected somewhere in between, such as 12/31/92. Generally, A is defined based on the data extraction protocol (i.e., from when the data is available) and C is defined by

25 the last day for which the member is still enrolled and eligible for the benefits. Of course, variations of those general points of definition could be selected within the scope of the present invention.

The definition of B is important. In the present invention, two basic definitions of B were devised in order to maximize the accuracy of the prediction model. Although, as would be understood by those skilled in the art, alternative definitions of B are contemplated.

30

Figure 6A is a first exemplary time window scheme, referred to as Scheme 1, for use in processing the data from the event level files shown in Figure 4.

In Scheme 1, the event prediction region is set from B to C such that  $B=C-(x\# \text{ of months})$  for all the members in the analysis. For example, if a 6-month depression hospitalization (HL) model (i.e., HL is used as a dependent variable) is to be built then  $B=C-(6 \text{ months})$ . In Jane Doe's example, B would equal 12/31/92.

35

Therefore, only data covering from A through B (1/1/91-12/31/92) is used to predict the depression HL in the 'next 6 months'. The phrase 'next 6 months' in this context implies that the time point B is "NOW" and any time after it is in the FUTURE and any time before it is in the PAST. This is a key concept of Scheme 1 and is important to understanding the prediction model implementation and application.

As additional explanation, when a variable is defined such as '# of Psychotherapy Visits in the LAST 6 Months', that means that the count for this variable is based on claims from [(B-6 months) to B] for every member in the analysis. It should be noted, however, that point B may vary with every member in the analysis population.

An alternative to Scheme 1, and referred to as Scheme 2, is illustrated in Figure 6B which shows a second exemplary time window scheme for use in processing the data from the event level files shown in Figure 4.

A difference between Scheme 1 and Scheme 2 is the definition of the prediction region for members which have at least one depression hospitalization or emergency room visit (HL/ER). The prediction region starting at point B, in Scheme 2, is defined in multiple passes over each member's record. Turning again to Jane Doe's analysis record (from 1/1/91 through 6/30/93, A=1/1/91, C=6/30/93) to illustrate how this aspect works for defining point B, assume that Jane Doe was hospitalized for depression three times: on 4/1/91, 4/1/92, and 4/1/93.

Point B is set equal to the date of the first depression HL/ER - 1 month or set equal to point C if a member never had depression HL/ER in their claims history. For Jane Doe, B=4/1/91. In the exemplary embodiment of the present invention, moving back one month from the HL date is performed to simulate the model application environment. There would probably be at least 30-day lag from model scoring to the disease management actions based on the scoring reports. Thus, in Jane Doe's record  $B=4/1/91-(1 \text{ month})=2/28/91$ . Jane's record, in this case, would not be used in the model building because the time span of the analysis region is only two months--less than the exemplary six month data history requirement.

Repeating steps 1 and 2 using second (or third or...) HL date to set point B, Jane Doe's record would eventually make it into model building on the second and third pass. This process, in the exemplary embodiment, terminates after three or four passes since there would probably be very few members with five or more depression HL/ERs in the study population.

It should be noted that the consequence of repeated modeling introduces added complexity of setting up additional independent variables. An important advantage, however, of Scheme 2 is that the prediction HL/ER rate would likely be higher than in Scheme 1.

5 In still another alternative embodiment, analysis weights which reflect proximity to the event to be predicted can be used, for example, within 3 months x 1, 3-6 months x .75, 6-9 months x .5, 9-12 months x .25, greater than 12 months x .125. Other suitable weighting techniques, as will be appreciated by those skilled in the art, could be used. These type of weighting techniques may  
10 be used with either Scheme 1 or Scheme 2.

Therefore, given a selected time window scheme and an appropriate set of predetermined variables, the processing step of 122 generates the analysis file.

Using the analysis file, the model for identification/prediction can then be developed in various ways using statistical techniques. In particular, the analysis  
15 file, now at a member level, is processed using statistical functions available in SAS. In the exemplary embodiment of the present invention, the statistical processing performed to generate the prediction model is multiple logistic regression. As will be appreciated by those skilled in the art, other statistical techniques may also be suitable for use with the present invention.

20 In the exemplary embodiment, the statistical processing, when applied to the analysis file, identifies variables which meet predetermined levels of significance (e.g., probability value < 0.05). These variables then form a prediction model which is a mathematical equation of the following form:

$$\text{Logit}(p) = a + bx_1 + cx_2 \dots + z x_i$$

25 where  $x_1 \dots x_i$  are the identified variables and  $a \dots z$  are their parameter estimates. An individual's probability (p) for the outcome under consideration is then determined using the following formula:

$$p = e^{-\text{logit}(p)} / (1 + e^{-\text{logit}(p)})$$

Figure 7A shows experimental results for a model based on Scheme 1 and  
30 using the HL indicator as a dependent variable. The resulting independent variables selected for the prediction model include "FEMALE INDICATOR", "# DEP HLS", "# DEP ERS", "# DR/PSYCHIATRIST VISITS", "# DEP PROCS", and "# OTH HLS AND DEP RELATED CODE".

Figure 7B shows experimental results, including the dependent variables,  
35 for a model also based on Scheme 1 but using, as the dependent variable the High Cost indicator.

It should be noted that, although both experimental results indicate that six independent variables were used for the prediction model, more or less independent variables could be used based on their individual ability to accurately predict the selected dependent variable.

5           Next, the determined model is applied to the data. That is, because the prediction zone in the above processing was actually based on past data for analysis purposes, the model is now applied to the data such that a prediction zone is defined in the future. The determined model can be applied to the existing data, to the data as it is regularly updated or to other claims databases for other  
10   benefits providers. To do so, only the determined independent variables of interest need to be processed. Of course, as new claims databases are to be analyzed, the entire process can be repeated to generate a new model in order to determine if other variables may be better predictors. The output generated by applying the model is a file containing a list of all of the depression patients and a  
15   indicator representative of the likelihood that that patient will have an adverse health outcome (i.e., experience that defined by the dependent variable). This list can then be divided into subgroups such as in 5% or 10% increments of patients likely to have the adverse health outcome.

          Applying the model to future claims data or other databases of depression  
20   patients or building a new model in a new database as described above, depression patients at high risk can be identified allowing for various types of intervention to maximize the effective allocation of health care resources for depression patients. Such intervention may take the form of 1) specific case management, 2) novel interventions based on subgroup characteristics, 3) high risk intervention, 4) high  
25   (relative) cost intervention, or 5) plan modification all adhering, of course, to the best practice guidelines.

Appendices I-VI follow.



## Appendix I

## Depression ICD-9-CM Codes

5

ICD-9-CM Code	Description
296.2x	major depressive disorder, single episode
296.3x	major depressive disorder, recurrent episode
296.5x	bipolar affective disorder, depressed
296.82	atypical depressive disorder
298.0x	depressive type psychosis
300.4x	neurotic depression (dysthymia)
311.xx	depressive disorder, not elsewhere classified

## Appendix II

## 5 CPT-4 or ICD-9-CM Procedure Codes for Psychotherapy

ICD-9-CM Procedure Codes:

10

9426	Sub convulsive electroshock therapy
9427	Other electroshock therapy
943-9439	Individual psychotherapy
44-9449	Psychotherapy and counseling

CPT-4 Codes:

908xx	All psychiatric procedure codes
90841-90844	Individual medical psychotherapy
90846-90849	Family medical psychotherapy
90855	Interactive individual psychotherapy
90857	Interactive group psychotherapy
90870-90871	Electroconvulsive therapy

### Appendix III

**Antidepressants Agents**  
**(From the DPS national formulary, 1995)**

**Therapeutic class 5.5.1.1 (Tertiary Amines)**

**amitriptyline**

10 doxepin

**imipramine**

**trimipramine**

**clomipramine**

**Therapeutic class 5.5.1.2 (Secondary Amines)**

15 desipramine

**nortriptyline****amoxapine**

**protriptyline**

**Therapeutic class 5.5.1.3 (Selective Serotonin Reuptake Inhibitors)**

20 paroxetine

**sertraline**

**fluoxetine**

**Therapeutic class 5.5.1.4 (Other Antidepressants)**

**amitriptyline/perphenazine**

25                    trazodone

burpropion

**venlafaxine**

**Therapeutic class 5.5.2 (Monoamine Oxidase Inhibitors)**

**isocarboxazid**

30 phenelzine

**tranylcypromine**

## Appendix IIIa

## Neuroactive drugs not for depression

35

all codes 5.x not in appendix above

## Appendix IV

Drugs possible used in excess by patients with severe depression  
DPS 1995 Formulary

5

## Possible related drugs

## 51. Analgesics and other medications for headache

10 9.1 Antacids

9.2 Antidiarrheal

9.3 Antispasmodic

9.4 Antiulcer

9.5 Laxiative

15 9.6 Other GI

11.1.1 Salicylates

11.1.2 Non-steroidal anti inflammatory drugs

11.3.1 Direct muscle relaxants

20 11.3.2 CNS muscle relaxant drugs

11.4 Other muscle relaxants

12.1.2 Multivitamins, fluorides, B2, Folic Acid, therapeutic vitamins

13.1.1 Prenatal vitamins

25 13.7 Oral contraceptives

15.2.1 Antihistamines

15.2.2 Decongestants

15.2.3 Combination antihistamines/decongestants

30

15.3 Antitussives and expctorants

Appendix V  
Major illness diagnosis

5

## ICD-9 code

10	Neoplasm (any site, any type)	140-239
	Ischaemic heart disease (any form)	410-414
	Pulmonary heart disease	415-417
	Heart Failure	428
	Cerebrovascular disease	430-438
15	Chronic obstructive pulmonary disease	490-496
	Non infectious enteritis and colitis	555-558
	Nephritis, nephrotic syndrome and nephrosis	580-589
	Normal delivery and other indication for care.....	650-659
	Injury and Poisoning	800-999
20	Suicide risk	300.9
	Attempted suicide- by drug	E9502-E952

Appendix VI

5

Major procedures

Essentially these will be considered as any surgical procedure

10 CPT code 10040-69979

15

Minor procedures

These are multiple screening tests and drug screening

20 CPT codes 80002-80103

What is claimed is:

1. A computer-implemented method for generating a model to identify at risk patients diagnosed with depression, information about patients existing in a claims database, said method comprising the steps of:

- 5 processing, based on predetermined criteria, the patient information in the claims database to extract claims information for a group of depression patients;  
defining, using the information available in the claims database, a set of events relevant to depression;  
creating, using the extracted claims information and the defined events,  
10 files containing event level information;  
defining a time window for providing a timeframe from which to judge whether events should be considered in subsequent processing;  
defining a set of variables as potential predictors;  
processing the event level information, using the time window and the set  
15 of variables, to generate an analysis file; and  
performing statistical analysis on the analysis file to generate a prediction model for use in identifying at risk patients diagnosed with depression, said prediction model being a function of a subset of the set of variables.

2. A computer-implemented method for identifying at risk patients  
20 diagnosed with depression, information about patients existing in a claims database, said method comprising the steps of:

- processing, based on predetermined criteria, the patient information in the claims database to find and extract claims information for a group of depression patients;  
25 defining, using the information available in the claims database, a set of events relevant to depression;  
processing the extracted claims information and the defined events to create files containing event level information;  
defining a time window for providing a timeframe from which to judge  
30 whether events should be considered in subsequent processing;  
defining a set of variables as potential predictors;  
processing the event level information, using the time window and the set of variables, to generate an analysis file;  
performing statistical analysis on the analysis file to generate a prediction  
35 model, said prediction model being a function of a subset of the set of variables;  
and

applying the prediction model to a processed claims database to identify and output a file listing the likelihood of each patient having an adverse health outcome.

3. The computer-implemented method of claim 1, wherein the step of  
5 processing extracts patients having been diagnosed with depression or prescribed an anti-depressant drug.

4. The computer-implemented method of claim 1, wherein the step of  
defining a set of variables includes defining both dependent and independent  
variables and a hospital (HL) indicator is defined as a dependent variable, where  
10 independent variables are representative of predictors and the dependent variable is representative of a adverse health outcome.

5. The computer-implemented method of claim 1, wherein the step of  
defining a set of variables includes defining both dependent and independent  
variables and a high cost indicator is defined as a dependent variable, where  
15 independent variables are representative of predictors and the dependent variable is representative of a adverse health outcome.

6. The computer-implemented method of claim 1, wherein the step of  
defining a set of variables includes defining both dependent and independent  
variables, substantially all of the data elements from the claims information as  
20 well as at least one combination of data elements are used as independent variables.

7. The computer-implemented method of claim 1, wherein the step of  
performing statistical analysis includes performing logistic regression.

8. An apparatus for generating a model to identify at risk patients  
25 with depression, information about patients existing in a claims database, said apparatus comprising:

means for processing, using predetermined criteria, the patient information  
in the claims database to find and extract claims information for a group of  
depression patients;

30 a predetermined set of events, derived from the claims information, said events being relevant to depression;

means, using the extracted claim information and set of events, for  
creating files of event level information;

a predetermined time window for providing a timeframe from which to  
35 judge whether events should be considered in subsequent processing;



a predetermined set of variables representing potential predictors;  
means, using the time window and the set of variables, for processing the event level information to generate an analysis file; and

means for performing statistical analysis on the analysis file to generate a  
5 prediction model used for identifying at risk patients diagnosed with depression, said prediction model being a function of a subset of the set of variables.

9. The apparatus of claim 8, further comprising:

means for applying the prediction model to a processed claims database to identify and output a likelihood for each patient of having an adverse health  
10 outcome.

10. A computer-readable medium containing a program for generating a model to identify at risk patients diagnosed with depression from a claims database which contains information about patients, said program on said medium comprising:

15 means for causing a computer to process, based on predetermined criteria, the patient information in the claims database to extract claims information for a group of depression patients;

means for causing the computer to input a set of predetermined events relevant to depression;

20 means for causing the computer to create, using the extracted claims information and the defined events, files containing event level information;

means for causing the computer to establish a time window for providing a timeframe from which to judge whether events should be considered in subsequent processing;

25 means for causing the computer to input a set of predetermined variables representative of potential predictors;

means for causing the computer to process the event level information, using the time window and the input set of variables, to generate an analysis file; and

30 means for causing the computer to perform statistical analysis on the analysis file to generate a prediction model used for identifying at risk patients diagnosed with depression, said prediction model being a function of a subset of the set of variables.

11. A computer-implemented method for identifying at least one patient  
35 with a specified disease who is at risk for a near-term high-cost clinical outcome associated with said disease and at least one clinical event predictive of at least one high-cost clinical outcome attributed to said patient, and providing early notice to a

medical services provider of said at-risk patient or said patient, and optionally providing to the provider or patient one or more interventions which can modify said clinical event and risk of high-cost clinical outcome, which method comprises:

- 5           i) collecting relevant data on patients who have a current diagnosis of the specified disease from one or more data sources;
- ii) entering or merging said data into a single electronic data file;
- iii) optionally cleaning up said data file by removing extraneous data and associating patient data with the correct patient identifier;
- 10          iv) identifying a clinically relevant, appropriate time period;
- v) creating a patient data file identifying, within the clinically relevant time period, for each patient the presence and frequency of at least one predetermined clinical event predictive of risk for a near-term high-cost clinical outcome relevant to the specified disease;
- 15          vi) stratifying the identified patients into risk groups based on the found predetermined clinical event by processing through a computer the patient data file, using a second pre-existing data file, wherein the second pre-determined data is created using a statistical procedure and a data file generated by the steps of:
  - 20           a) processing, based on predetermined criteria, patient information in at least one clinically relevant data source to extract information for a group of patients with the specified disease;
  - b) defining, using the information available in said data source, a set of potentially predictive events relevant to a high cost clinical outcome in patients with the specified disease;
  - 25           c) defining a time-window from which to judge whether events should be considered in subsequent processing;
  - d) creating files for predicting clinical outcomes containing relevant data rearranged to represent the set of potentially predictive events; and
  - 30           e) running a statistical procedure to identify a set of predictive events which in combination show a statistically significant association with a high-cost clinical outcome;
  - vii) notifying the healthcare provider of said patient or patient that said outcome is likely to occur in the near term, and
  - 35           viii) optionally providing to the health care provider or patient, interventions which modify the high-cost clinical outcome or reduce its severity, which interventions are optionally derived from the one or more of the set of

predictive clinical events which in combination show a statistically significant association with a high-cost clinical outcome.

12. The method of claim 11 wherein the patients have been diagnosed with depression or prescribed an anti-depressant drug.

5 13. The apparatus of claim 11 wherein the patient data is drawn from at least one data source which is a pharmacy, hospital, or physician records data source.

14. An apparatus for generating a data set identifying at least one patient with a specified disease who is at risk for a near-term high-cost clinical  
10 outcome associated with said disease and at least one clinical event predictive of at least one high-cost clinical outcome attributed to said patient, and providing early notice to a medical services provider of said at-risk patient or said patient, and optionally providing to the provider or patient one or more interventions which can modify said clinical event and risk of high-cost clinical outcome, said apparatus  
15 comprising:

i) means for collecting relevant data on patients who have a current diagnosis of the specified disease from one or more data sources;

ii) means for entering or merging said data into a single electronic data file;

20 iii) means for optionally cleaning up said data file by removing extraneous data and associating patient data with the correct patient identifier;

iv) means processing the electronic data file to identify a clinically relevant, appropriate time period;

25 v) means for creating a patient data file to identify, within a pre-determined clinically relevant time period, for each patient the presence and frequency of at least one predetermined clinical event predictive of risk for a near-term high-cost clinical outcome relevant to the specified disease;

30 vi) means for stratifying the identified patients into risk groups based on the found predetermined clinical event by using a processing means wherein the patients data file is processed against a second pre-existing data file, wherein the second pre-existing predictive data file is created by an apparatus comprising:

a) means for processing, based on predetermined criteria, patient information in at least one clinically relevant data source to extract information for a group of patients with the specified disease;

35 b) means for defining, using the information available in said data source, a set of potentially predictive events relevant to a high cost clinical outcome in patients with the specified disease;

c) a pre-determined time-window from which to judge whether events should be considered in subsequent processing;

d) means for creating files for predicting clinical outcomes containing relevant data rearranged to represent the set of potentially predictive events; and

e) means for running a statistical procedure to identify a set of predictive events which in combination show a statistically significant association with a high-cost clinical outcome.

15           15.     The apparatus of claim 14 wherein the specified disease is  
10     depression.

          16.     The apparatus of claim 14 wherein the patient data is drawn from at least one data source which is a pharmacy, hospital, or physician records data source.

          17.     A computer-readable medium for controlling a computer and  
15     containing a program for generating a data set identifying at least one patient with a specified disease who is at risk for a near-term high-cost clinical outcome associated with said disease and at least one clinical event predictive of at least one high-cost clinical outcome attributed to said patient, and providing early notice to a medical services provider of said at-risk patient or said patient, and optionally  
20     providing to the provider or patient one or more interventions which can modify said clinical event and risk of high-cost clinical outcome, said program on said medium comprising:

          i) means for causing a computer to receive and store in a data file relevant data on patients who have a current diagnosis of the specified disease from one or  
25     more types of electronic input;

          iii) means for causing a computer to rearrange or delete certain data in the data file to remove extraneous data and to associate patient data with the correct patient identifier;

          iv) means for causing the computer to establish a time window for  
30     providing a timeframe for selecting data in the data file which falls into a clinically relevant, appropriate time period;

          v) means for causing a computer to create a second patient data file from the first data file by causing the computer to identify, within the pre-set clinically relevant time period, from the first data file those patients which have at least one  
35     predetermined clinical event predictive of risk for a near-term high-cost clinical outcome relevant to the specified disease and frequency of that clinical event for each selected patient;

vi) means for causing a computer to stratify the patients in the second data file into risk groups based on the found predetermined clinical event by causing the computer to process the second patient data file against a third pre-existing data file, wherein the third pre-existing predictive data file was created by a

5 program on a computer-readable medium comprising:

a) means for causing a computer to process, based on predetermined criteria, patient information in at least one clinically relevant data source to extract information for a group of patients with the specified disease;

10 b) means for causing the computer to input a set of potentially predictive events relevant to a high cost clinical outcome in patients with the specified disease;

c) means for causing the computer to establish a time window for providing a timeframe from which to judge whether events should be  
15 considered in subsequent processing;

d) means for causing the computer to create files for predicting clinical outcomes containing relevant data rearranged to represent the set of potentially predictive events; and

e) means for causing the computer to perform a statistical analysis  
20 to identify a set of predictive events which in combination show a statistically significant association with a high-cost clinical outcome;

vii) means for causing the computer to output the data file of the stratified patient groupings; and optionally

viii) means for causing a computer to generate and output a set of  
25 interventions tailored to informing patients identified as being at high-risk of a near-term high-cost clinical outcome of said risk and means for preventing or reducing said risk.

18. The computer-readable medium of claim 17 wherein the specified disease is depression.

30 19. The computer-readable medium of claim 18 wherein the patient data is drawn from at least one data source which is a pharmacy, hospital, or physician records data source.

20. A method for manufacturing one or more set of materials providing for interventions in the health management of patients at risk of a near-term high-cost clinical outcome, which method comprises identifying at least one patient  
35 with a specified disease who is at risk for a near-term high-cost clinical outcome associated with said disease and at least one clinical event predictive of at least one

high-cost clinical outcome attributed to said patient, and providing early notice to a medical services provider of said at-risk patient or said patient, and optionally providing to the provider or patient one or more interventions which can modify said clinical event and risk of high-cost clinical outcome, said program on said

5 medium comprising:

i) means for causing a computer to receive and store in a data file relevant data on patients who have a current diagnosis of the specified disease from one or more types of electronic input;

10 iii) means for causing a computer to rearrange or delete certain data in the data file to remove extraneous data and to associate patient data with the correct patient identifier;

iv) means for causing the computer to establish a time window for providing a timeframe for selecting data in the data file which falls into a clinically relevant, appropriate time period;

15 v) means for causing a computer to create a second patient data file from the first data file by causing the computer to identify, within the pre-set clinically relevant time period, from the first data file those patients which have at least one predetermined clinical event predictive of risk for a near-term high-cost clinical outcome relevant to the specified disease and frequency of that clinical event for  
20 each selected patient;

vi) means for causing a computer to stratify the patients in the second data file into risk groups based on the found predetermined clinical event by causing the computer to process the second patient data file against a third pre-existing data file, wherein the third pre-existing predictive data file was created by a  
25 program on a computer-readable medium comprising:

a) means for causing a computer to process, based on predetermined criteria, patient information in at least one clinically relevant data source to extract information for a group of patients with the specified disease;

30 b) means for causing the computer to input a set of potentially predictive events relevant to a high cost clinical outcome in patients with the specified disease;

c) means for causing the computer to establish a time window for providing a timeframe from which to judge whether events should be  
35 considered in subsequent processing;

d) means for causing the computer to create files for predicting clinical outcomes containing relevant data rearranged to represent the set of potentially predictive events; and

5 e) means for causing the computer to perform a statistical analysis to identify a set of predictive events which in combination show a statistically significant association with a high-cost clinical outcome;

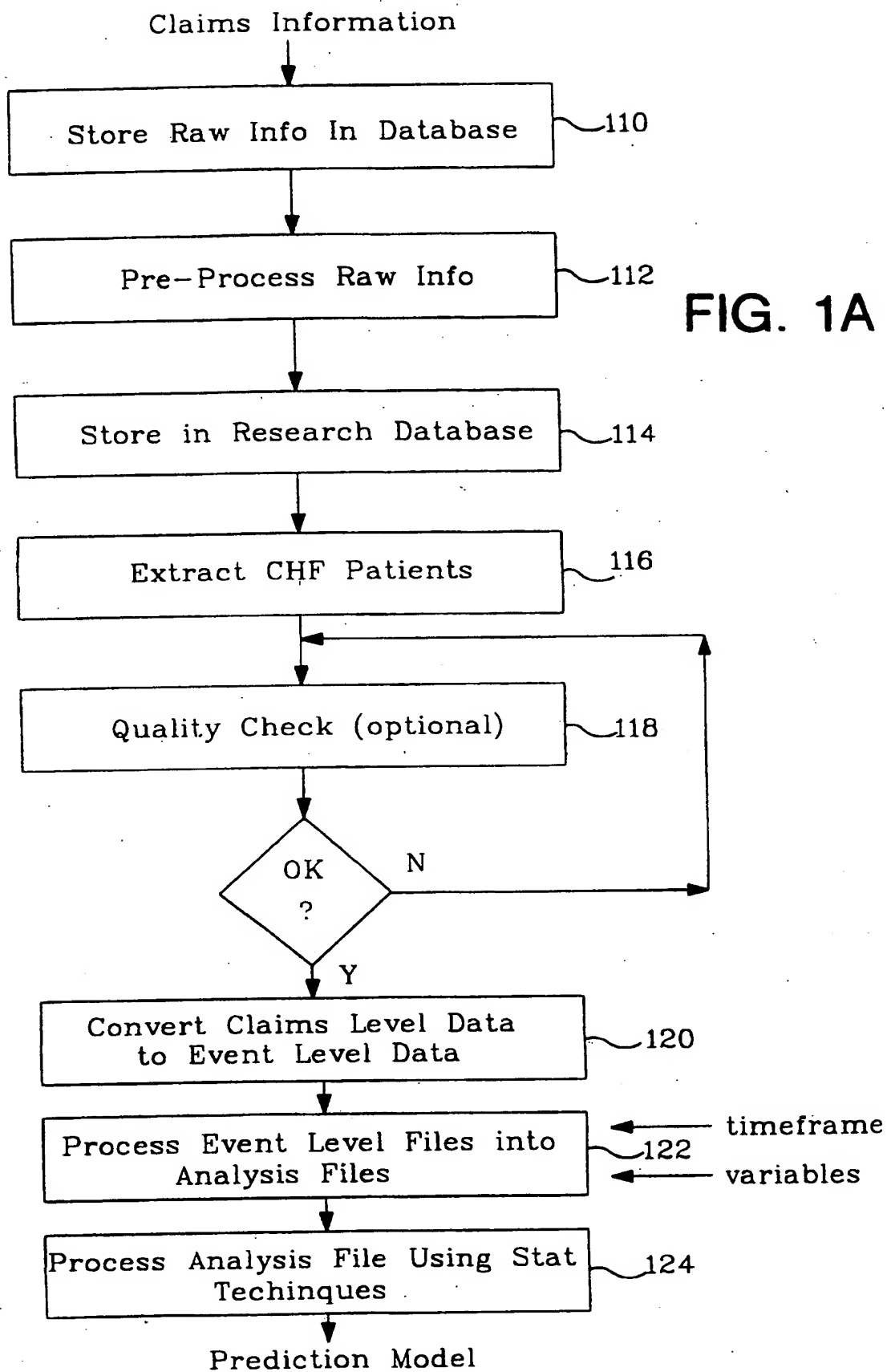
vii) means for causing the computer to output the data file of the stratified patient groupings; and

10 viii) means for causing a computer to generate and output a set of interventions tailored to informing patients identified as being at high-risk of a near-term high-cost clinical outcome of said risk and means for preventing or reducing said risk.

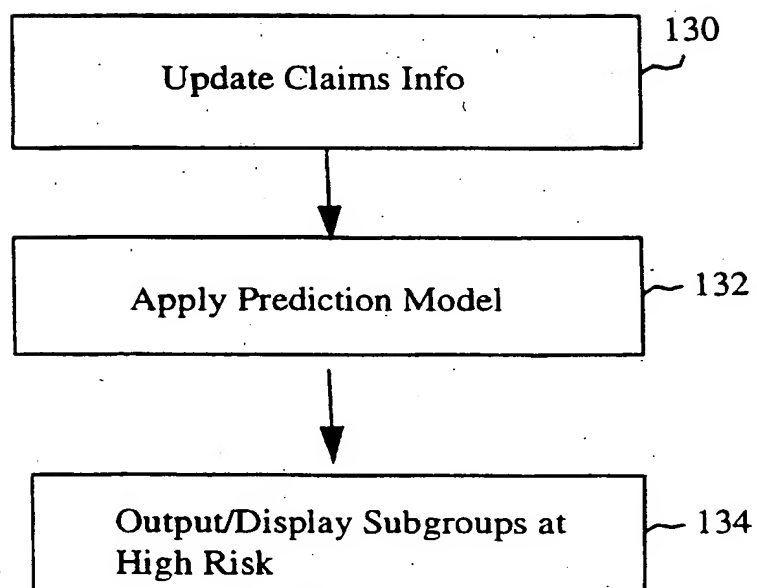
21. The article of claim 20 wherein the product is specific to depression.

15 22. The article of claim 20 wherein the product is based on patient data drawn from at least one data source which is a pharmacy, hospital, or physician records data source.

1 / 10





**FIG. 1B**

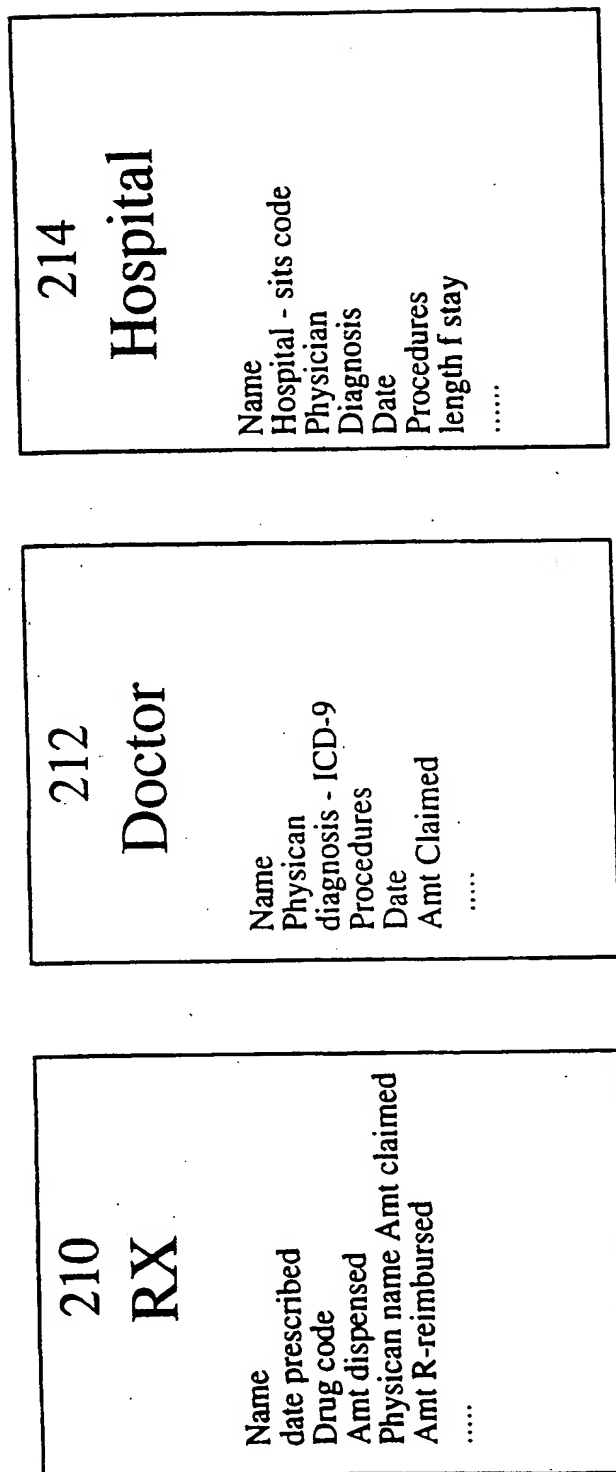


FIG. 2

SAS Format

Rx	DR	HL
ID, date drug...	ID, ICD, date...	ID, ICD, hosp...
claim 1		
*		
*		
*		
claim x		

FIG. 3

Event		Event Date		Number	Cost	Category Indicator
1	Hospitalization for depression	date		LOS	\$	severity
2	Emergency room for depression	date		blank	\$	blank
3	Doctor (outpatient) visit for depression	date		blank	\$	specialist
4	Prescription for SSRI	date		days supply	\$	therapy class
5	Prescription for TCA	date		days supply	\$	therapy class
6	Prescription for other neuroactive drug	date		days supply	\$	sub-class
7	Procedure for depression	date		blank	\$	sub-class
8	Hospitalization not for depression	date		LOS	\$	severity
9	Emergency room not for depression	date of first		number in month	\$	
10	Doctor (outpatient) visit not for depression	date of first		number in month	\$	
11	Prescription for possibly related drugs	date of first		number in month	\$	
12	Prescription for non-depression drugs	date of first		number in month	\$	
13	Procedure not for depression	date of first		number in month	\$	Severity indicator

FIG. 4

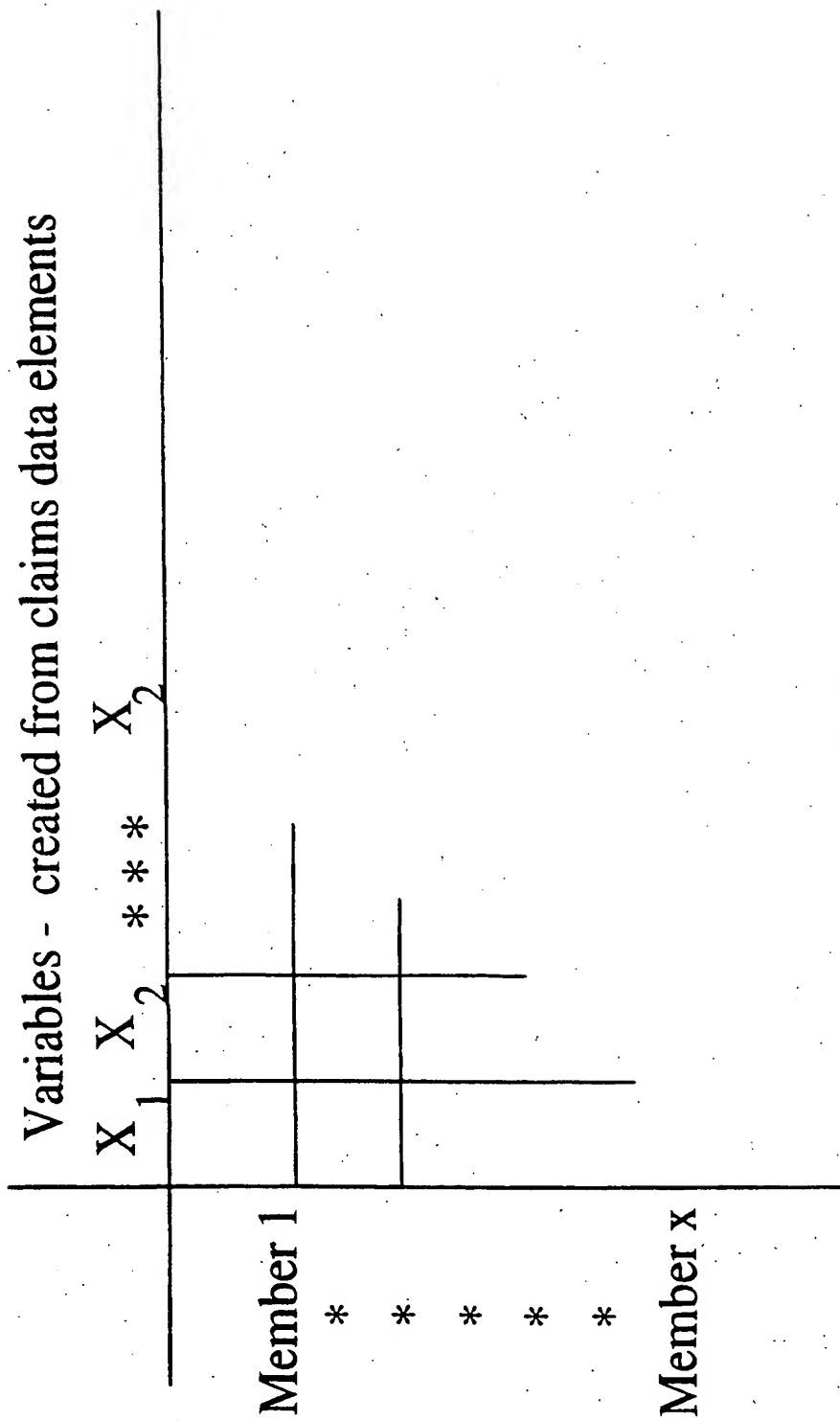


FIG. 5

7/10

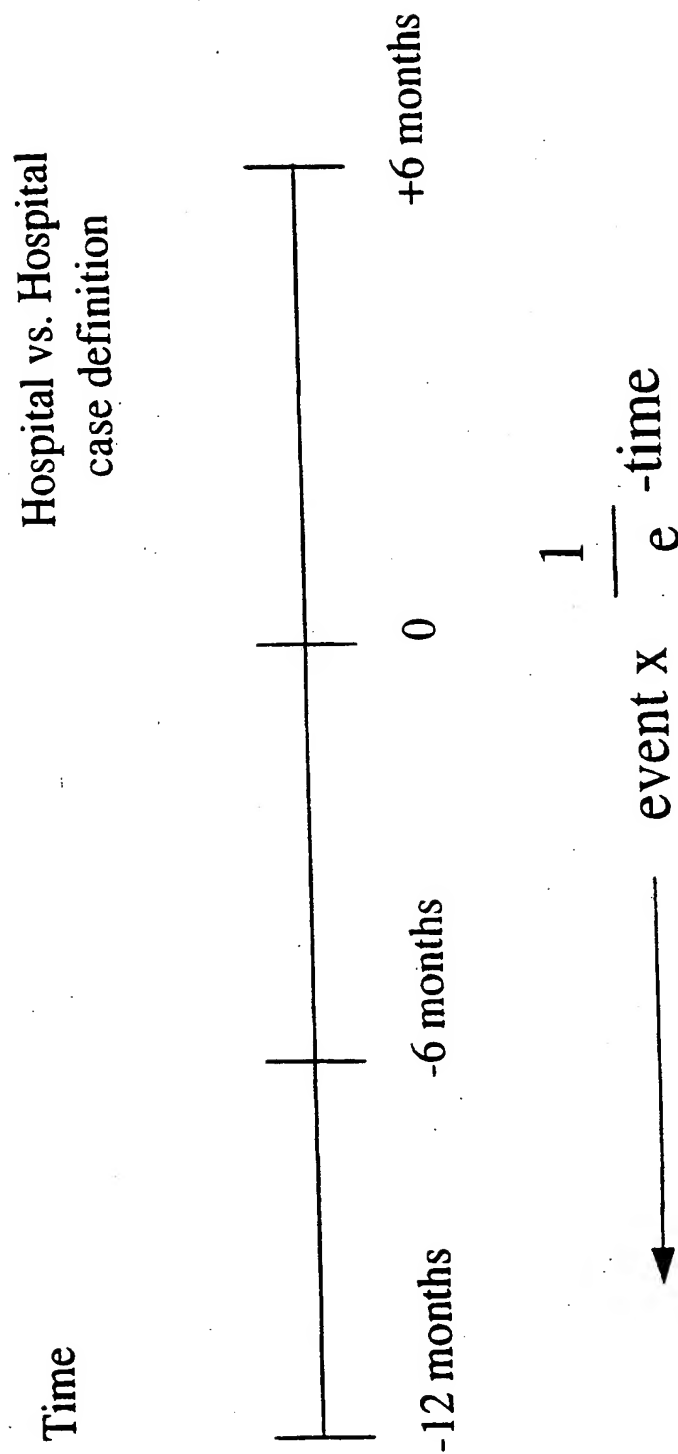


FIG. 6A

Case definition  
Last Hospital

Time

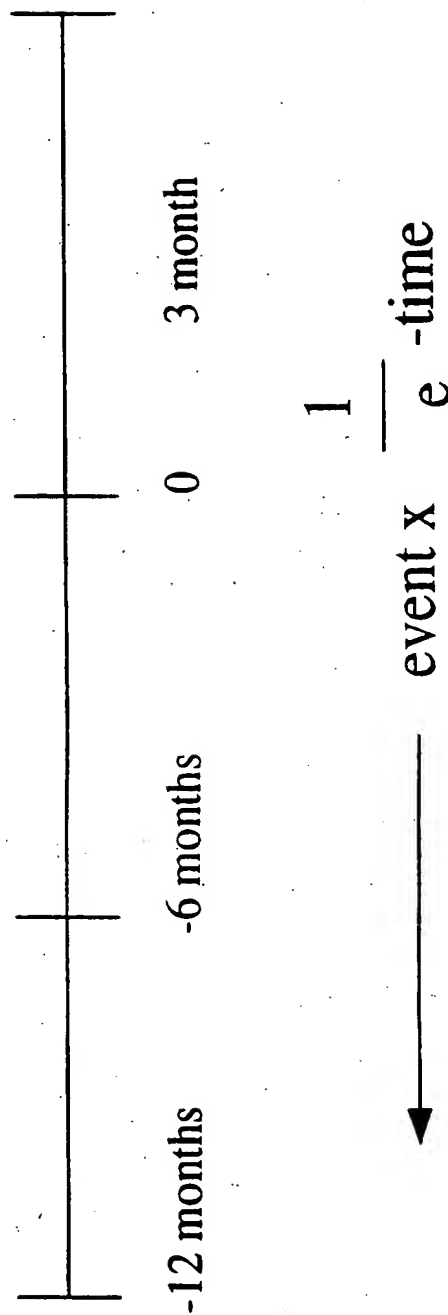


FIG. 6B

9 / 10

Variable	parameter estimate	Odds ratio (CI)	P (chi square)
Intercept	-4.6010	0.010 (.006-.016)	1.0 x 10 <sup>00</sup>
Depression Hosp	1.2401	3.456 (2.26-5.286)	1.0 x 10 <sup>-8</sup>
Psychiatric Hosp	1.3942	4.032 (1.99-8.155)	1.0 x 10 <sup>-4</sup>
ER (non-dep)	0.5719	1.772 (1.33-2.36)	9.0 x 10 <sup>-5</sup>
Depression Proc	0.0627	1.065 (1.03-1.09)	1.0 x 10 <sup>-5</sup>
Number Gen Drs	1.0651	2.901 (1.74-4.85)	5.0 x 10 <sup>-5</sup>
Female	0.7037	0.495 (0.326-0.75)	9.0 x 10 <sup>-4</sup>

**FIG. 7A**



10 / 10

Variable	parameter estimate	Odds ratio (CI)	P (chi square)
Intercept	-2.3632	0.090 (0.0376-.237)	$5.0 \times 10^{-7}$
Sedative/anxio.	0.1100	1.116 (1.092-1.141)	$1.0 \times 10^{-5}$
OPD Psychiatrist	0.083	1.087 (1.044-1.131)	$5.0 \times 10^{-5}$
SSRI	0.1955	1.216 (1.165-1.270)	$1.0 \times 10^{-5}$
Psychotherapy	0.0437	1.065 (1.012-1.079)	$7.6 \times 10^{-3}$
Number Gen Drs	0.9159	2.499 (1.87-3.34)	$6.0 \times 10^{-10}$
Female	0.4695	0.625 (0.476-0.822)	$8.0 \times 10^{-4}$

FIG. 7B

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US97/01829

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G01N 33/48; A61B 5/00; G06F 17/60

US CL : 395/204, 201, 202, 203; 424/9, 11; 128/630, 898; 283/117, 900

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 395/204, 201, 202, 203; 424/9, 11; 128/630, 898; 283/117, 900  
G06F 17/60; G01N 33/48; A61B 5/00;

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, Image Search of above subclasses

search terms: risk, risk#, analysi?, process?, insurance, predict?, forecast?, estimat?, model?, simulat?

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P	US 5,594,637 A (EISENBERG et al) 14 January 1997	1-22
A	US 4,740,364 A (HODGEN) 26 April 1988	1-22
A	US 5,396,886 A (CUYPERS) 14 March 1995	1-22

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:	* T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
* A* document defining the general state of the art which is not considered to be part of particular relevance	* X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
* E* earlier document published on or after the international filing date	* Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
* L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	* G*	document member of the same patent family
* O* document referring to an oral disclosure, use, exhibition or other means		
* P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

25 MARCH 1997

Date of mailing of the international search report

17 APR 1997

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

GAIL O. HAYES

Telephone No. (703) 305-9711

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☐ BLACK BORDERS

☒ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

☐ FADED TEXT OR DRAWING

☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING

☐ SKEWED/SLANTED IMAGES

☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS

☐ GRAY SCALE DOCUMENTS

☒ LINES OR MARKS ON ORIGINAL DOCUMENT

☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**

**THIS PAGE BLANK (USPTO)**